Augmented World Models

Philip Ball*, Cong Lu*, Jack Parker-Holder, Stephen Roberts

UNIVERSITY OF OXFORD

Summary of Main Contributions

- Dynamics augmentation for offline RL, allowing us to be robust to changing dynamics **training only on a single setting**.
- Propose a simple **self-supervised context adaptation** algorithm, significantly increasing zero-shot performance.
- Both approaches offer **significant improvement** v.s. SotA methods.

Offline reinforcement learning

We work with model-based Dyna-style offline RL. But what if the test environment differs from the training environment?

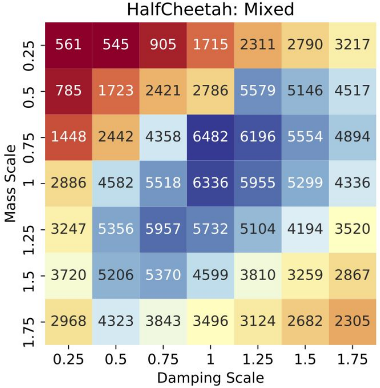
Training Data

Billions of available samples

Test Environment

New data expensive to collect

MOPO

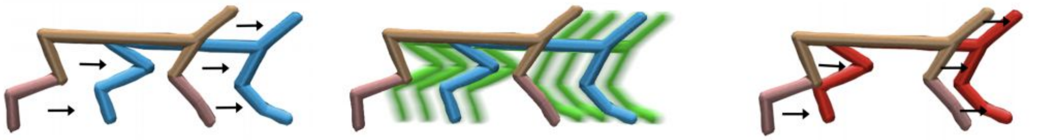


Question: Can we generalize with just the training data?

Can we have our (Le) Cake and eat it?

Problem: We only have one set of dynamics in our offline dataset.

Idea: Augment the dynamics, to produce a variety of random settings for our agent to train on.

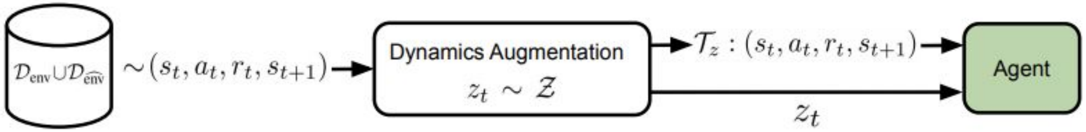


(a) World Model: $\widehat{P}$ (b) Augmented World Model: $\widehat{P}_z$ (c) Test: $P^\star$

The augmentation that worked best was Dynamics Amplitude Scaling (**DAS**):

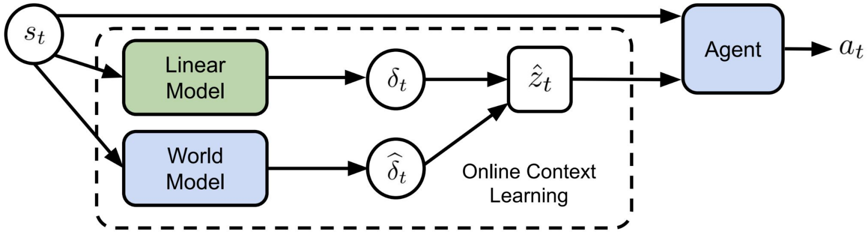
$$\mathcal{T}_z : (s_t, a_t, r_t, s_{t+1}) \mapsto (s_t, a_t, r_t, s_t + z \odot \delta_t)$$

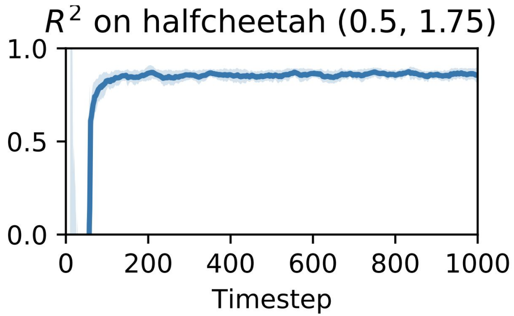
The **AugWM** training approach:



Augmentations & Linear Context Learning

This augmentation is passed to the policy as a context, and then recovered via a linear model at test-time!





Can AugWM improve zero-shot generalisation?

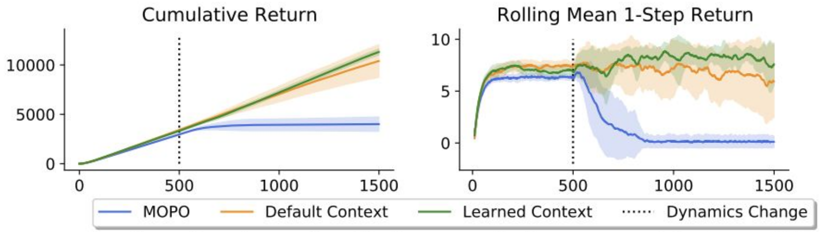
Standard D4RL with mass/damping

Dataset Type	Environment	MOPO	AugWM (Ours)
Random	HalfCheetah	2303 ± 112	$2818 \pm 197 \star$
Random	Walker2d	569 ± 103	706 ± 139
Mixed	HalfCheetah	3447 ± 218	$3948 \pm 122 \star$
Mixed	Walker2d	946 ± 95	$1317 \pm 206 \star$
Medium	HalfCheetah	2954 ± 89	2967 ± 106
Medium	Walker2d	1477 ± 337	1614 ± 440
Med-Expert	HalfCheetah	1590 ± 766	$2885 \pm 432 \star$
Med-Expert	Walker2d	1062 ± 334	$2521 \pm 316 \star$

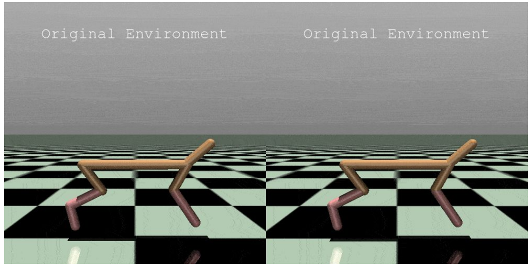
More complex environments

Setting	MOPO	AugWM (Default)	AugWM (LM)
Ant: Mass/Damp	1634	1715	**1804**
Ant: One Crippled Leg	1370	1572	**1680**
Ant: Two Crippled Legs	700	697	**795**
HalfCheetah: Big	4891	**5194**	4968
HalfCheetah: Small	5151	**5488**	5263

Adaptation during an episode



AugWM agents are more *robust* and can *adapt* to changes in test-time dynamics



Left = MOPO, Right = AugWM.

Discussion

- We present Augmented World Models (AugWM), which:
  - Introduces the offline -> online changed dynamics problem
  - Makes the agent more robust
  - Improves zero-shot generalization
- In future:
  - Investigate meta-learning for few-shot learning
  - Additional non-stationary test-time environments
  - Pixel-based tasks with latent states